DOCUMENT RESUME

ED 368 154 FL 021 399

AUTHOR Bachman, Lyle F.; And Others
TITLE Investigating Variability in Tasks and Rater Judgments in a Performance Test of Foreign Language Speaking.
PUB DATE Aug 93
NOTE 27p.; Paper presented at the Annual Language Testing Research Colloquium (15th, Cambridge, England, August 2-4, 1993).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Students; Higher Education; *Interrater Reliability; *Language Proficiency; *Language Tests; Second Language Learning; Spanish; Speech Skills; *Test Reliability; *Test Theory
IDENTIFIERS University of California

ABSTRACT

This paper outlines the development of a performance assessment measure of language speaking ability, the Language Ability Assessment System (LAAS), which is highly reliable and can be examined for reliability through modern measurement theories, such as generalizability theory (G-theory) and the many-facet Rasch theory. LAAS was developed to determine which University of California students were ready for full academic immersion in a foreign country whose language of instruction is not English. Test takers read passages and view recorded lectures in the target language, answering questions in writing and orally, and summarizing the lectures orally. LAAS was tested on 218 University of California students planning to spend an academic year abroad in a Spanish-speaking country. An analysis of the scores, raters, and the test itself found that, according to G-theory measures, the reliability of all the scales, with the possible exception of pronunciation, were well within acceptable limits. Many-facet Rasch measurements demonstrated how individual raters and tasks could affect the estimation of a test-taker's language ability, finding a large range of severity in the raters' judgments of performance. (MDM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 368 154

Investigating Variability in Tasks and Rater Judgments in a Performance Test of Foreign Language Speaking

Lyle F. Bachman
Univdersity of California, Los Angeles

Brian K. Lynch
University of Melbourne

Maureen Mason
University of California, Los Angeles

FL021399

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Lyle F. Bachman
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Investigating Variability in Tasks and Rater Judgments in a Performance Test of Foreign Language Speaking

Lyle F. Bachman
Brian K. Lynch
Maureen Mason

INTRODUCTION[1]

Much of the recent debate that has surrounded the development and use of "performance", or "communicative" language tests has focused on a supposed trade-off between two sets of desirable qualities. It has been argued by proponents of such tests that test tasks and test performance need to correspond to non-test language use. This suggests that we should design tests with characteristics such as thematic unity and task dependency, which seem to be features of non-test discourse, and that we should present test takers with authentic tasks that require them to interpret and create authentic language. At the same time, we realize that the primary purpose of language tests is to assess, rather than to teach, or to provide opportunities to learn, so that in addition to incorporating features of language use into test design, we must demonstrate that scores derived from test performance are reliable, and that the inferences about language ability that we make from them are valid.

One area that has been of particular concern with performance tests is the potential variability in tasks and rater judgments, as sources of measurement error (e.g., Shohamy 1983, 1984; Pollitt & Hutchinson 1987, Barnwell 1989, Cumming 1990, McNamara & Adams 1991). This variability has been investigated in the language testing literature with two complementary approaches: generalizability theory (e.g., Bolus et al. 1982, Stansfield & Kenyon 1991) and many-facet Rasch modeling (e.g., McNamara & Adams 1991). Generalizability theory (Cronbach et al 1972; Brennan 1983; Shavelson & Webb 1991) provides a methodological approach to estimating the relative effects of variation in test tasks and rater judgments on test scores. Many-facet Rasch measurement (Linacre 1989, 1993) estimates differences in measurement facets such as task difficulty and rater severity, and adjusts ability estimates of test takers, taking these differences into account.

Generalizability theory

Generalizability theory (G-theory) constitutes both a theoretical framework and a set of procedures for specifying and estimating the relative effects of different factors on test scores. G-theory conceptualizes a person's performance on a language test as being a function of several different factors, or *facets*, in addition to the person's language ability. In a test consisting of several different items, for example, the facet of concern would be that associated with differential item difficulty. In a test that includes different tasks and different raters, the two facets of interest are those associated with differential difficulty of tasks and differential severity of raters. In tests with two facets, we also need to be concerned with the interactions between persons and the facets. If some tasks are differentially difficult for different groups of test takers, this may be a source of bias. Similarly, if raters score the performance of different groups of test takers differently, then this could be an indication of rater bias. Interactions such as these cannot be examined by reliability estimates, such as internal consistency and inter-rater consistency, that are derived from classical test theory. Since interactions between persons and tasks, persons and raters and between raters and tasks are frequently the major sources of error variance, traditional approaches to estimating inter-task and inter-rater consistency, or reliability are clearly inadequate.

In order to use G-theory to estimate the affects of the facets of concern to us for a given testing situation, we conduct a generalizability study (G-study) in which the effects of these different facets are clearly distinguishable, and which provides sample statistics indicating the magnitude of these effects in the sample used in the G-study. These G-study sample statistics provide the basis for a D-study, in which we estimate population values for the different sources of variation in the D-study design. The D-study can provide us with two types of information. First, we can estimate the relative importance of the effects associated with the different facets, and the interactions among the facets, in the form of *variance components*. These variance components provide estimates of the relative effects on an individual's test score of: 1) his or her ability, 2) the relative difficulty of the tasks, 3) the relative severity of the raters, and 4) the interactions between persons and tasks, persons and raters, and tasks and raters. Second, we can estimate the dependability, or reliability, of our test scores, taking into account the effects of the different facets and interactions. Estimates of dependability that are appropriate for both norm-referenced and criterion-referenced interpretations can be obtained.

Many-facet Rasch measurement

Many-facet Rasch measurement represents an extension of the one-parameter Rasch model, one of several models developed within Item Response Theory (IRT). These IRT models conceptualize a person's expected performance on a test item or task as a function of their *ability* and characteristics of the test task, such as the *difficulty* of the item, or task, the task's capacity for *discriminating* between high and low scorers, and the effect of *guessing*. The Rasch model is a one-parameter model because it uses only one task characteristic--task difficulty--in the estimation of person ability. Alternatively, person ability and task difficulty may be considered as *facets* in the measurement process, as is done in the many-facet extension of the Rasch model (Stahl & Lunz 1992). When the measurement is obtained using a rating scale of some sort, many-facet Rasch measurement can make use of the additional facet, rater severity, to refine the estimation of the test taker's ability. In many-facet Rasch analysis, therefore, separate estimates of task difficulty and rater severity can be obtained, and these, in turn, are used to estimate each test taker's ability. That is, if a particular rater is unusually lenient in her ratings, this would be taken into account when estimating the ability of a test taker who had been rated by that rater. Similarly, if a particular task is relatively difficult, this would be taken into account in estimating the ability of all test takers who responded to that task. This provides the potential for estimating the abilities of different individuals on the same scale of measurement, even though they may have been rated by different raters and may have responded to different tasks.

Purpose of the paper

In this paper we will attempt to demonstrate the following:

1) that it is possible, even feasible, to develop a "performance" assessment of speaking that is highly reliable and dependable, from both NR and CR perspectives,
2) that tools of modern measurement theory, such as G-theory and many-facet Rasch measurement, are essential for proper estimation of reliability in performance assessment, and

3) G-theory and many-facet Rasch measurement are not antithetical, but can be used together to provide valuable complementary types of information for the test development process.

In this paper we discuss the design and development of an innovative foreign language test battery that was designed for placing undergraduate applicants to the University of California Education Abroad Program into academic programs in universities outside of the United States that are appropriate for their level of language ability. We discuss the assumptions about language and the design principles that underlie the test battery and describe the battery itself. We then present the results of our investigation into task and rater variability in the grammar ratings of the speaking subtest of the battery, based on the performance of an operational administration to a group of University of California undergraduate students who had been selected for participation in the Education Abroad Program. In this investigation we utilized both G-theory and many-facet Rasch measurement. We conclude by discussing the implications of these results for the use of G-theory and many-facet Rasch measurement for the development of performance tests of language ability.

METHODOLOGY

Instrumentation: Language Ability Assessment System (LAAS)

The Language Ability Assessment System (LAAS) was developed to meet a growing need to determine which University of California students were ready for full academic immersion in a foreign country whose language of instruction is not English. The purpose of the test is two-fold: first, to place students into the appropriate academic program, either full academic immersion or a sort of "sheltered language" program, and second, to provide examinees with diagnostic feedback, in particular so those students who are not yet ready for full academic immersion can have a better idea of what specific areas of the language they should work on.

This project provided us an opportunity to empirically investigate and refine our knowledge about the nature of language ability, language use, and performance on language tests, while at the same time developing an assessment procedure that would be useful in a real-world situation. The design of the LAAS incorporates assumptions and principles that we believe are consistent with recent research in applied linguistics, as well as with current thinking in language testing. As these assumptions and principles are also consistent with what we know about language learning and language teaching, we would

hope that this test will reinforce a positive interface between instruction and assessment. These assumptions and principles are illustrated in Figure 1.

FIGURE 1 ABOUT HERE

The Structure of the LAAS

These guiding principles are reflected in the structure of the test, which is illustrated in Figure 2.

FIGURE 2 ABOUT HERE

The instructions for the LAAS are all in English, while the tasks are presented in the foreign language. For Part 1 of the LAAS, test takers read a passage taken from an academic text written in the foreign language and answer a series of open-ended reading comprehension questions. In Part 2, they first view an introductory academic lecture given in sheltered language to prepare them for the authentic academic lecture. Next, they view the academic lecture, which is a 10-12 minute segment taken from a lecture given by a native-speaking lecturer, normally before a class of native-speaking students at a university in the foreign country. Test-takers are encouraged to take notes during both lectures. Following the lectures, the test-takers answer a series of open-ended listening comprehension questions, which are presented in writing and to which they respond in writing. The topic of the reading passage and both of the lectures is the same. For Part 3, test-takers are asked to produce two speech samples, speaking into a tape recorder. Finally, in Part 4, they write an essay in which they are asked to synthesize the information presented in the reading passage and the lectures, and to relate it to something in their major field of study or their personal life.

In addition to being based on the theory that language ability is not measurable by one global factor, the design of the LAAS enables us to provide the students with detailed diagnostic information regarding their language ability. Scores are reported to the students in the form of a language ability profile, which includes scores for

listening and reading, ratings for five components each for speaking and writing, and composite scores for speaking and writing.

Speaking tasks

The speaking tasks are presented in the form of a role play. Test-takers are asked to imagine that they are going to visit the professor who presented the first lecture, during his office hours. For the first task the professor asks them to summarize the lectures. They have one minute to prepare their summary and three minutes to speak it into the tape recorder. They are also encouraged to ask the professor questions if they finish their summary before their time is up. For the second task the test takers are asked to relate a theme or concept from the academic lecture to their own personal or academic experience. For this task they also have one minute to prepare their response and three minutes to speak it into the tape recorder.

Administrative Procedures

The LAAS is administered entirely by video tape, and takes approximately two hours to complete. The test is typically administered in a language laboratory equipped with individual booths and tape recorders and a large video monitor.

Scoring Procedures

The speech samples were scored on criterion-referenced, analytic scales that measure five components of language ability: pronunciation, vocabulary, cohesion, organization, and grammar. The first four of these components are measured on a 1-4 scale with 1 being "no evidence of skill", 2 "poor", 3 "moderate", and 4 "good"[2]. On these scales, a rating of 3, "moderate", was set as the criterion level indicating readiness for a full academic program abroad. There is no 5 on the scale since the purpose of the test is to see which students are ready for full immersion and not to discriminate among those who are thus identified as ready. The grammar component was rated on a 1-7 scale, with 1 indicating "no systematic evidence" and 7 indicating "complete range with no systematic errors". On the grammar scale, a rating of 4, "large, but not complete range, with many error types", was set as the criterion level indicating readiness for a full academic program abroad.

Since we focus on the grammar ratings in this paper, we provide the grammar scale descriptors in Figure 3.

FIGURE 3 ABOUT HERE

Each of the two speech samples (summary, discussing theme from the lecture) was rated independently by at least two different raters. In cases where there was discrepancy between the first two raters across the criterion level ("3" on pronunciation, vocabulary, cohesion and organization; "4" on grammar), a third rater rated the speech samples. The two closest ratings were used for calculating the score. Four ratings--two for each of the two tasks--for each of the five scales, were averaged to arrive at scores for each test taker.

The raters were graduate students and faculty of the Department of TESL and Applied Linguistics and the Department of Spanish and Portuguese at UCLA. All were native or near-native Spanish speakers. Raters were all given a program of training, consisting of an orientation to the test, viewing the lecture and speaking prompts, a discussion of the rating scales, listening to, rating and discussing several sample tapes, followed by a norming session. Operational ratings were done in a single day, with periodic norming after breaks, in the language laboratory at UCLA.

Subjects

The subjects of the LAAS administration were regularly matriculated University of California students at eight University of California campuses. Most of the students were in the sophomore year of their undergraduate education, and all had been admitted to the academic year abroad program in a Spanish-speaking country; approximately 75% of the students were going to Spain, and 25% to Mexico. 218 students completed the speaking part of the LAAS, which is the focus of this study. The LAAS was administered at all eight campuses within a two-week period in March 1993.

RESULTS

G-Theory

Dependability of All Scales

Because we did not obtain three ratings operationally for all test takers, we did not have a fully balanced design utilizing three ratings. Operationally, we analyzed the scores obtained with the two closest ratings, irrespective of whether these were arrived at from the first two ratings or from two of three ratings. In addition, in order to obtain results that would be based on exactly the same data as the FACETS analyses, we compiled a data set that included only the first two raters, rather than the closest two, as was done operationally. In both data sets, the two ratings given to each of the two tasks were not consistently associated with particular raters, since G-theory assumes that ratings and raters are drawn from a relatively homogenous universe, and are thus randomly parallel. The G-study design used in this study was thus a fully-crossed design with two random facets--ratings and tasks--each with two conditions.[3] All analyses were conducted using GENOVA (Crick & Brennan 1983), a program for conducting generalizability studies. For comparison, the generalizability coefficients (ρ^2) and dependability coefficients (Φ) for these two sets of ratings are given in Table 1 below.

TABLE 1 ABOUT HERE

From Table 1 we can see that with the possible exception of pronunciation, the reliabilities of all the scales are well within acceptable limits. We can also see that the gains in reliability associated with using third ratings to resolve discrepancies were minimal, the largest difference being only .037 for cohesion. However, because of the need to obtain scores that would be clearly either above or below the a priori criterion ability levels set, third ratings will continue to be used to resolve discrepancies across the criterion levels.

Grammar Ratings[4]

In interpreting the results of the grammar ratings, our first interest was to look at the relative sizes of the D-study variance components associated with persons, ratings, tasks, and the interactions among these, since these would tell us what the major sources of measurement error in our model would be. These variance components for the D-study with two ratings and two tasks are given in Table 2 below.

TABLE 2 ABOUT HERE

The largest variance component was that associated with persons, which accounts for 91 per cent of the total variability in these measurements. This can be interpreted as universe, or "true" score variance, or the amount of variance that can be reliably associated with differences in test taker's levels of ability. Looking at the main effects for ratings and tasks, we can see that these are both negligible, indicating that the ratings were not different in severity, and that the tasks were of about equal difficulty for all test takers. Of the variance components associated with interactions, that for the person by rating interaction (pR) accounts for the second largest proportion of the total variance (six per cent). This interaction indicates a very slight tendency for ratings to systematically vary across different groups of test takers. This suggests that there may be some sub groupings of either raters or test takers that are behaving differently from the rest of their groups, and that this may be a potential source of rater bias. The relatively small person by task interaction (pT) indicates that the two tasks are not differentially difficult for different groups of test takers. Finally, the negligible rater by task interaction (RT) indicates that raters are not rating the two tasks differentially.

The next area of interest was the dependability of the scores as indicators of grammatical competence. Since we designed this as a criterion-referenced test, we looked at the domain-referenced dependability index, Φ, which is .914. This can be interpreted analogously to a norm-referenced reliability coefficient, and means that about 91% of the observed score variance is universe score variance. Another useful index of dependability is the signal-to-noise (S/N) ratio, which tells us how much more universe score variance there is than error score variance. In our data, the S/N ratio is

10.63, which means that there is roughly ten times as much universe score variance as error variance in these scores.

We can also estimate the dependability of mastery/non-mastery decisions based on different cut scores. One estimate of this is the squared-error loss agreement index, Φ_λ, where λ is the cut score. Agreement indices and S/N ratios for three different possible cut scores, 4, 5 and 6, are given in Table 3 below.

TABLE 3 ABOUT HERE

The agreement index at the a priori cut-score of 4 (.922) indicates that in addition to providing dependable measures of grammatical competence, the a priori cut-scores are providing dependable information for making placement decisions.

In addition to obtaining estimates of variance components and dependability for the actual numbers of conditions of the facets in the G-study, we can estimate these for different combinations of conditions. That is, we can estimate what our dependability indices would be if we had a greater or lesser number of ratings or tasks. Criterion-referenced dependability estimates and S/N ratios for different combinations of numbers of ratings and tasks are given in Table 4 below.

TABLE 4 ABOUT HERE

From these results we can see that sizable increases in dependability are appreciated by including two tasks, rather than one. They also demonstrate that only minimal increases would result from obtaining three ratings routinely for all tests. However, as the results of Table 1 show, the focused third ratings that are currently being done operationally result in higher reliabilities.

Many-Facet Rasch Measurement

Reliability

The many-facet Rasch analogs of the classical reliability estimate, or more appropriately in this case, of the G-theory generalizability coefficient, are the separation and reliability indices. The person separation indices produced by FACETS are the familiar Rasch statistics that indicate "the spread of the estimates relative to their precision" (Linacre & Wright 1993, p. 66), while the person reliability indices are the "Rasch equivalent of the KR-20 or Cronbach Alpha statistic" (Linacre & Wright 1993, p. 66). These indices for all five rating scales of the speaking part of the LAAS are given in Table 5 below.

TABLE 5 ABOUT HERE

Specific facet effects

In contrast to generalizability theory, which provides information that is in effect averaged across the elements in each facet, many-facet Rasch measurement is able to provide information concerning individual persons, raters, tasks, and elements of any other facets specified in the measurement model. This information is provided on a common scale, the log-linear scale which has *logits* as the basic unit of measurement. Rather than raw scores, then, many-facet Rasch reports person ability, task difficulty, rater severity, and other facets that may be included in the design of the measure, as probabilities. The logit scale relates probabilities associated with person ability to probabilities associated with task difficulty, and other facets. Furthermore, these probabilities are arrived at by considering all the information available from the measurement process--the raters are ranked in relation to how likely they are to give particular persons particular scores on particular tasks; the tasks are ranked in relation to how likely they are to be associated with particular raters giving particular scores to particular persons, and so forth. The FACETS program iteratively relates the information on each facet to the others and comes up with estimates for each person's ability, each rater's severity, and each task's difficulty, all expressed on the same, probabilistic scale. In addition, through a procedure called *bias analysis* (the analysis

of interactions), it is possible to identify particular raters or particular tasks that are having an unusual effect on the estimation of person ability. Thus, while generalizability theory can identify *interactions* between sources of test variability, many-facet Rasch measurement can identify the individual sources of those interactions, for example, particular rater-task combinations, or rater-person combinations. All many facet analyses were conducted on the LAAS data with FACETS Version 2.68 (Linacre & Wright, 1993). For the purposes of this study, we will focus on the reliability of the ratings as measures of persons, the relative severity of raters, the relative difficulty of tasks, and the bias (interaction) analysis for persons by raters, persons by tasks, and raters by tasks.

Relative severity of different raters

As indicated above, FACETS provides estimates of rater severity that are on the same scale, the logit scale, as are the other facets. The relative severity of the 15 raters, as estimated by FACETS, is presented in Table 6 below.

TABLE 6 ABOUT HERE

The 15 raters demonstrated a relatively large range of severity in their judgments, with rater 17 being the most severe (logit value = 1.93) and rater 2 being the least severe (logit value = -2.27). This large range is also reflected in the relatively large Separation Index (3.31) and Reliability Index (.92). In addition to looking at the relative severity of the raters, we can examine whether or not each individual rater "fit" the measurement model; that is, whether or not the individual raters are consistent within their own ratings. To do this, we look at the Infit Standardized Mean Square (*Infit MnSq Std*) value. If the value of this statistic is greater than or equal to the absolute value of 2, the rater is considered as **not** fitting the model. Only Raters 5 and 12 were identified by this criterion; however, the **negative** standardized values are not as serious as a positive misfit, since this indicates that the rater simply had too little variation in his or her ratings. FACETS also provides *model error*, which indicate the standard errors of the estimates of rater severity. As can be seen, these model errors are all quite small, indicating that the estimates are relatively stable.

These results may appear to be at odds with the GENOVA results, which indicated an estimate of zero for the variance component for ratings. suggesting that the ratings were not different in terms of their overall severity. However, the FACETS results indicate that the large range in severity is accounted for primarily by two extreme raters, Raters 17 and 2, and that on average, the raters do not differ that much in their severity.

Relative difficulty of different tasks

As with rater severity, FACETS provides estimates of task difficulty, also on the logit scale. Table 7 presents the results of the FACETS analysis for **tasks**.

TABLE 7 ABOUT HERE

Although the logit values for the two tasks were different--0.24 for Task 1 and -0.24 for Task 2, the absolute difference is relatively small, as are the Separation Index (2.20) and Reliability Index (.83), suggesting that the tasks were relatively similar in difficulty. However, Task 2 has an Infit MnSq Std of -2, which indicates less variance in this task than would have been expected from the model.

Analysis of interactions

The analysis of interactions, or bias analysis, as it is called by FACETS, looks at pairs of facets and reports the number of *inconsistent* ratings, or ratings that are random in there departure from the pattern identified in the overall analysis, and the number of *biased* ratings, or ratings that demonstrate a consistent pattern that is different from that identified for the overall analysis. Table 8 displays the number of significantly (Z-scores of absolute values equal to or greater than 2) inconsistent and biased ratings, along with the number of ratings used for estimating the bias terms, for each of the three two-way interactions involving persons, raters, and tasks in the LAAS data.

TABLE 8 ABOUT HERE

As can be seen from Table 8, there were very few ratings that were either significantly inconsistent or biased. The largest proportion, seven out of 394, was for the person by rater interaction, and in this case these involved only four persons and five raters, since some of the persons and raters were involved in more than one interaction.

These results again parallel those from the GENOVA analysis, where the person by rater interaction was associated with the largest variance component (other than that for the object of measurement, persons). The person by task interaction was relatively small, as it was in the FACETS analysis. Both analyses make the important point that, while there may be consistency among the ratings resulting in what would be considered more than acceptable inter-rater reliability, in terms of classical measurement theory, there can be evidence of systematic differences in the way the ratings are being given that needs to be investigated. The FACETS bias analysis allows us to identify specific person-rater combinations that are not being properly accounted for in the analysis, and gives us information about whether the rating for the particular person was more severe or less than expected given the overall pattern of ratings.

DISCUSSION

For many years language testers have been interested in assessment methods that correspond to non-test language use situations, and this interest has been heightened by the broadened view of language use and language ability that has informed much recent discussion in language testing. These so-called "performance" assessment methods typically present test takers with tasks that are more complex and extensive than the stand-alone items that have characterized much of modern language testing since the early 1960's. Furthermore, in such assessment an additional dimension, or facet, the rater, is added to the procedure. Because of the difficulty, indeed, the undesirability, of rigidly controlling the nature of the tasks, and because of the subjectivity of rater judgments, variations across tasks and raters, as potential sources of measurement error, have been of concern to proponents and critics of performance testing, alike.

In this paper we believe we have demonstrated that a language assessment procedure that involves several complex tasks and in which test takers' performance is rated by multiple raters can yield scores that are consistent across both tasks and raters. The results of two approaches to estimating reliability--G-theory and many-facet Rasch

measurement--both indicate the scores obtained from this assessment procedure are reliable at separating test takers in terms of their language ability, and are highly dependable indicators of the domain of language ability they are aimed at measuring. Furthermore, they provide dependable information for making criterion-referenced placement decisions about test takers.

We believe that we have also shown that in order to obtain reasonable estimates of reliability in performance assessment, we must utilize measurement models that permit us to investigate multiple sources of measurement error simultaneously. Generalizability theory and many-facet Rasch measurement provide us with tools to estimate the magnitude of multiple sources of variability in scores obtained from performance tests that include multiple tasks and multiple raters. This, in turn, allows us to estimate the dependability of our measures with greater precision than when using traditional measurement theory, which can estimate only a single source of variation at a time.

It is our experience that G-theory and many-facet Rasch measurement, rather than being antithetical models of measurement, can provide complementary information that is useful to test developers and test users. In addition to improving our global estimates of the dependability of test scores, G-theory and many-facet Rasch measurement allow us to identify specific elements of our testing and research procedures that are affecting those scores. G-theory identifies the relative effects of facets such as raters or tasks, as well as the relative effects of combinations of these facets (interactions). Many-facet Rasch measurement allows us to identify specific raters, specific tasks, and specific combinations of raters, tasks, and persons that are affecting the dependability of our judgments.

In test development, we can proceed as follows. Once we have identified the sources of variability in scores from language performance tests, G-theory allows us to estimate how much more dependable our judgments would be if we modified our assessment procedure; for example, by increasing the number of raters or tasks. Many-facet Rasch measurement allows us to accommodate differences in variability across facets such as raters and tasks when estimating person ability from language performance data. When we use the logit values for person ability, instead of raw scores, these differences in variability are taken into account. In addition, we can use the results of the bias (interaction) analysis to identify consistent deviations from the

pattern accounted for by the many-facet Rasch measurement, which can then be individually adjusted by hand. This analysis also allows us to more systematically investigate the effects of training raters and using particular tasks when collecting language performance data. Finally, information about individual rater's patterns of rating that are produced by FACETS can be used to provide feedback to raters in the hope of improving their ratings in the future.

NOTES

1 An earlier version of this paper was presented at the 15th Annual Language Testing Research Colloquium in Cambridge, England, 2-4 August 1993. We would like to thank Dr. Tim McNamara, of Melbourne University, for his assistance in understanding many-faceted Rasch measurement and for his comments on an earlier draft of this paper. We would also like to thank José Ramon Nuñez for his assistance in developing the prompts, and Amie Williams for her assistance in producing the video tape, for the speaking part of the LAAS.

2 These capsule descriptors are supplemented with more elaborate descriptors in the actual rating scales that are used. For example, the "moderate" (3) descriptor for vocabulary in Spanish reads as follows: "3 MODERATE: Vocabulary of moderate size. **Range**: Some variety of word choice; **Accuracy**: some Spanish words may be inaccurately used; occasionally relies on false cognates and/or English words; **Communication**: frequently misses or searches for Spanish words."

3 Because two specific tasks were used in the speaking part of the LAAS, we considered the possibility of treating task as a fixed facet. The disadvantage of this is that we would not then obtain estimates for all the possible variance components in the design. We therefore ran the data both ways, treating task as both a fixed and a random facet. The results indicated very little difference in the absolute values of the estimates that were obtained, and no difference in their relative magnitudes. We therefore decided to treat task as a random facet in order to obtain the full information on interactions, for potential diagnostic purposes.

4 Recall that the analyses of the grammar ratings are based on the data set with the first two ratings, rather than with the operational ratings.

REFERENCES

Barnwell, D. 1989. 'Naive' native speakers and judgments of oral proficiency in Spanish. *Language Testing, 6*, 152-163.

Brennan, R.L. 1983. *Elements of generalizability theory*. Iowa City, Iowa: The American College Testing Program.

Crick, J.E. and Brennan, R.L. 1983. *Manual for GENOVA: A GENeralized ANalysis of VAriance system*. Iowa City: The American College Testing Program.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rataratnam, N. 1972. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Cumming, A. 1990. Expertise in evaluating second language compositions. *Language Testing, 7*, 31-51.

Ellis, R. 1987. *Second Language Acquisition in Context*. London: Prentice-Hall.

Linacre, J.M. and Wright, B.D. 1993. *A User's Guide to FACETS: Rasch-model computer program, version 2.4 for PC-compatible computers*. Chicago, IL: MESA Press.

Linacre, J.M. 1989, 1993. *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

McNamara, T. 1990. Item Response Theory and the validation of an ESP test for professionals. *Language Testing, 7*, 52-76.

McNamara, T. and R. J. Adams. 1991. Exploring inter-rater reliability with Rasch techniques. Paper presented at the 13th Annual Language Testing Research Colloquium. Princeton: Educational Testing Service, March 1991.

Pollitt, A. and Hutchinson, C. 1987. Calibrated graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, 4*, 72-92.

Shavelson, R. J. and Webb, N. M. 1991. *Generalizability Theory: A Primer*. Newbury Park, CA: Sage.

Shohamy, E. 1983. The stability of oral proficiency assessment on oral interview testing procedure. *Language Learning 33*, 527-40.

Shohamy, E. 1984. Does testing method make a difference? The case of reading comprehension. *Language Testing 1*,147-70.

Stahl, J.A. and Lunz, M.E. 1992. A comparison of generalizability theory and multi-faceted Rasch measurement. Paper presented at the Midwest Objective Measurement Seminar, University of Chicago, May 1992.

Stansfield, C. W. and D. M. Kenyon 1992. Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. System 20, 3:347-364

Figure 1

Assumptions and Principles that Underlie LAAS Development

Assumptions

- Language ability is multicomponential
- Language use involves a dynamic interaction between language users, the context, and the discourse that evolves in a speech event;
- In order for scores from LAAS to be useful for its intended purposes (i.e., placement, diagnosis), performance on LAAS needs to be related to language use in academic settings abroad.

Principles

- Measure as many components of language ability as are relevant and feasible for the intended purposes.
- Provide test scores that reflect profiles of language ability
- Present test takers with authentic materials and tasks.
- Include content that is relevant to the academic contexts in which EAP students will be studying, and that will provide thematic unity to all parts of the test.
- Present test takers with tasks that require them to use language.

Figure 2.

Structure of the Test

Part 1: Reading
- Academic reading passage
- Reading comprehension questions

Part 2: Listening
- Introductory academic lecture
- Authentic academic lecture
- Listening comprehension questions

Part 3: Speaking
- Task 1: Summary of lectures
- Task 2: Extended response

Part 4: Writing
- Academic essay

Figure 3

Grammar Rating Scale

	RANGE		ACCURACY
1	No systematic evidence	AND	control of few or no structures; errors of all or most possible are frequent.
2	Limited range, but with some systematic evidence	AND	control of few or no structures; errors of all or most possible are frequent.
3	Limited range, but with some systematic evidence	AND	control of some structures used, but with many error types.
4	Large, but not complete range	AND	control of some structures used, but with many error types.
5	Large, but not complete range	AND	control of most structures used, with few error types.
6	Complete range	AND	control of most structures used, with few error types.
7	Complete range	AND	No systematic errors, just lapses.

Table 1

Generalizability Coefficients (ρ^2) and
Dependability Coefficients (Φ) for All Scales:
Operational Ratings and First Two Ratings

Scale	Operational Ratings	First Two Ratings
	ρ^2/Φ	ρ^2/Φ
Pronunciation	.768/.759	.752/.744
Vocabulary	.912/.911	.895/.893
Cohesion	.917/.915	.881/.878
Organization	.879/.877	.861/.859
Grammar	.935/.933	.916/.914
All Scales	.943/.941	.918/.916

Table 2

D-Study Variance Components: Grammar Scale

Effect	Variance component	Standard error	Percentage of total variance
p	1.718	0.180	91%
R	0.000	0.000	0%
T	0.003	0.003	0%
pR	0.107	0.015	6%
pT	0.011	0.006	1%
RT	0.000	0.000	0%
pRT,e	0.041	0.004	2%
Total	1.880		

Table 3

Agreement Indices for Different Cut Scores

Cut Score	Φ_λ	S/N Ratio
4	.926	12.48
5	.922	11.78
6	.959	23.47

Table 4

Dependability Estimates for Different Numbers of Ratings and Tasks

# Ratings	# Tasks	Φ	S/N Ratio
1	1	.810	4.250
2	1	.888	7.957
3	1	.918	11.223
1	2	.847	5.554
2	2	.914	10.635
3	2	.939	15.301

Table 5

Person Separation and
Reliability Indices for All Scales

Scale	Persons Separation/Reliability
Pronunciation	1.92/.79
Vocabulary	1.50/.69
Cohesion	2.26/.84
Organization	2.13/.82
Grammar	4.24/.95
All Scales	6.41/.98

Table 6

Relative Severity of Raters

Measure Logit	Model Error	Infit MnSq Std	Rater	
1.93	0.28	-1	17	more severe
1.18	0.30	1	13	
1.14	0.25	-2	5	
0.76	0.26	1	14	
0.47	0.26	-1	6	
0.30	0.24	0	8	
-0.10	0.28	0	11	
-0.12	0.30	0	16	
-0.38	0.29	-1	9	
-0.39	0.26	0	10	
-0.39	0.33	-2	12	
-0.50	0.39	-1	18	
-0.76	0.23	-1	4	
-0.86	0.23	-1	15	
-2.27	0.31	0	2	more lenient

Table 7

Relative Difficulty of Tasks

Measure Logit	Model Error	Infit MnSq Std	Task	
0.24	0.10	0	1	more difficult
-0.24	0.10	-2	2	less difficult

Table 8

Analysis of Interactions

Interactions	Person x Rater	Person x Task	Rater x Task
Inconsistent Ratings	0/790[1]	2/792	2/790
Biased Ratings	7/394[2]	3/396	0/30

[1] For the inconsistent ratings, the first number is the number of ratings that are inconsistent, while the second number is "the number of ratings used in estimating the bias terms. This may be only a small fraction of the entire data set." (Linacre & Wright 1993, p. 70).

[2] For the biased ratings, the first number is the number of ratings that are biased, while the second number is "the number of modelled bias terms in found in the data." (Linacre & Wright 1993, p. 72)